doi:10.1093/nar/gkr799

MACiE: exploring the diversity of biochemical reactions

Gemma L. Holliday[1,*], Claudia Andreini[2,3], Julia D. Fischer[1], Syed Asad Rahman[1], Daniel E. Almonacid[4], Sophie T. Williams[1] and William R. Pearson[5]

[1]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, [2]Magnetic Resonance Center (CERM), University of Florence, Via L. Sacconi 6, [3]Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy, [4]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA 94158 and [5]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908, USA

Received August 19, 2011; Revised September 7, 2011; Accepted September 12, 2011

ABSTRACT

MACiE (which stands for Mechanism, Annotation and Classification in Enzymes) is a database of enzyme reaction mechanisms, and can be accessed from http://www.ebi.ac.uk/thornton-srv/databases/MACiE/. This article presents the release of Version 3 of MACiE, which not only extends the dataset to 335 entries, covering 182 of the EC sub-subclasses with a crystal structure available (~90%), but also incorporates greater chemical and structural detail. This version of MACiE represents a shift in emphasis for new entries, from non-homologous representatives covering EC reaction space to enzymes with mechanisms of interest to our users and collaborators with a view to exploring the chemical diversity of life. We present new tools for exploring the data in MACiE and comparing entries as well as new analyses of the data and new searches, many of which can now be accessed via dedicated Perl scripts.

INTRODUCTION

Enzymes make the wonderful diversity of life possible, from thermophiles that exist under incredibly harsh conditions to the complexity of higher organisms, such as humans. However, despite their importance and our continued fascination with these often complex proteins we still have a relatively limited understanding of how they function. Since 1964, when the Enzyme Commission (EC) first published their rules for enzyme nomenclature and their system to classify the overall reaction that an enzyme performs (1), there have been over 5000 EC numbers assigned, although 836 have been subsequently either transferred to other EC numbers, or deleted (data correct as of June 2011). The first proteins with a fully defined sequence and assigned identifier from the curated portion of UniprotKB (Swiss-Prot) (2) were deposited in the 1980s, and the first crystal structures relating to an enzyme were deposited in the wwPDB (3) in the early 1970s. Since then, the growth of information has been persistent (Figure 1A); however, there are still some significant gaps in our knowledge (Figure 1B).

Of the 4528 currently active EC numbers, only 2792 have a sequence in Swiss-Prot that has a fully assigned EC number (i.e. a catalytic activity with all four levels of the EC number assigned), and of those only 1761 also have an associated structure deposited in the wwPDB, although not all of these will have a reliable mechanism published in the primary literature. Despite this apparent lack of data, there is a great deal of knowledge available, including structures, gene sequences, mechanisms, metabolic pathways and kinetic data. However, these data tend to be spread between many different databases and throughout the literature. Most web resources relating to enzymes [such as BRENDA (4), KEGG (5), SABIO-RK (6), the IUBMB Enzyme Nomenclature website (1) and IntEnz (7)] focus on the overall reaction, accompanied in some cases by a textual or graphical description of the mechanism. MACiE (8,9), which stands for Mechanism, Annotation and Classification in Enzymes, is a collaboration between the Thornton group (EMBL-EBI), Mitchell group (University of St Andrews, Scotland) and Bertini group (University of Florence, Italy) and was designed to provide a computational description of mechanism by including detailed stepwise mechanistic information for a wide coverage of both chemical space and the protein structure universe. First published in 2005 (9),

[*]To whom correspondence should be addressed. Tel: 01223 492535; Fax: 01223 494486; Email: gemma@ebi.ac.uk

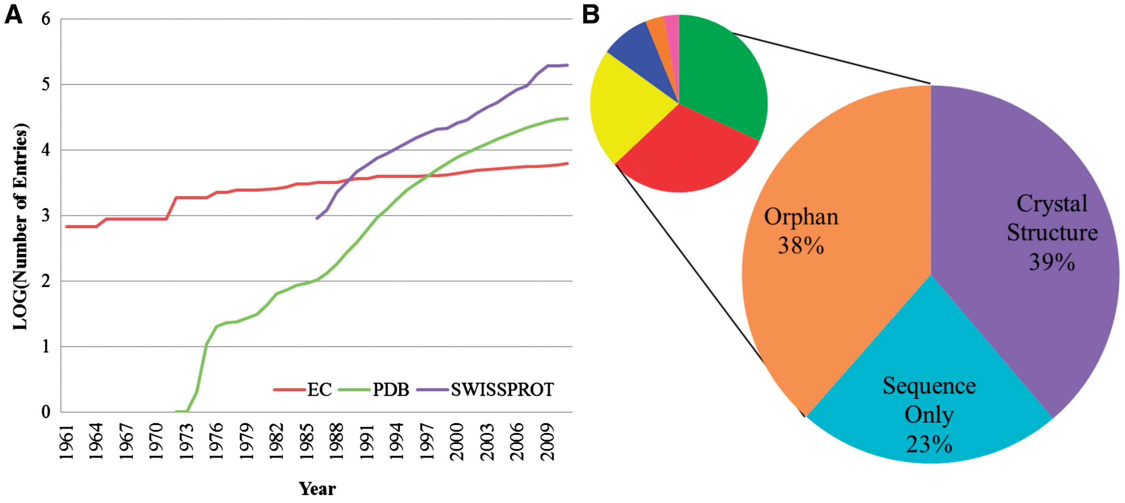


Figure 1. (**A**) The growth of biological data according to the EC (red), wwPDB (green) and Swiss-Prot (blue) databases (up to June 2011). (**B**) The large pie-chart shows the percentage of EC numbers covered by the wwPDB (purple) and Swiss-Prot (light blue), the inset (small pie-chart) represents the percentage breakdown of the orphan enzymes (those with no sequence or structure) by EC class (the oxidoreductases (EC 1) in green, the transferases (EC 2) in red, the hydrolases (EC 3) in yellow, the lyases (EC 4) in blue, the isomerases (EC 5) in orange and the ligases (EC 6) in magenta).

MACiE usefully complements both the mechanistic detail of the Structure–Function Linkage Database (SFLD) (10) which provides information for a small number of rather 'promiscuous' enzyme superfamilies, and the wider coverage with less chemical detail provided by EzCatDB (11) and the Catalytic Site Atlas (CSA) (12). Entries in MACiE are linked, where appropriate, to all of these related data resources. MACiE is also proving a useful resource for understanding how enzymes catalyse the vast array of chemistry with such a (relatively) limited repertoire of catalytic entities (13–16).

This new release of MACiE retains all the original features of previous releases, but includes enriched data content through the extension of data entries (next section), new tools for exploring the diversity of biochemical reactions in MACiE ('New Methods for Characterizing and Comparing Enzyme Mechanisms' section) as well as new searches and database statistics (see Supplementary Data). Each biologically meaningful search allows the user to not only access the individual entries, but also view the data in a comparative overview page. Many of these are now available as separate links and visualization of the database online has also been updated ('Updates to MACiE Website' section).

DATA CONTENT AND NEW ANNOTATIONS IN MACiE

This release of MACiE represents the addition of 133 new entries since the previous major release (bringing the total number of entries to 335). We now cover >90% (182) of the EC sub-subclasses with an available crystal structure, representing 321 distinct EC numbers. When we include related enzymes as defined using the distant homology described in the CSA, MACiE covers over 800 distinct EC numbers and over 17 000 PDB codes; with a stricter definition, statistically significant similarity using SSEARCH, an implementation of Smith-Waterman, MACiE covers over 600 EC numbers and over 7000 PDB codes. We have also incorporated new annotations, which will be described in the following subsections. With the incorporation of many homologues and functional analogues into MACiE, we have constructed some pre-defined datasets for users interested in specific aspects of MACiE, including datasets relating to the EC classification, diversity in structure and function, mechanistic diversity and other aspects such as cofactor requirements. For more detail on these, please see the Supplementary Data.

Cofactors in MACiE

In previous releases of MACiE (8,9), cofactor annotation was largely neglected. This has now been addressed, and there are two basic types of cofactors which are annotated in MACiE: metal and organic cofactors. Metal cofactors are primarily handled by Metal–MACiE (17), a sister database and collaboration with the Bertini group at CERM in Florence, Italy. Approximately half of all the entries in MACiE contain at least one metal ion (182 MACiE entries, covering 178 distinct EC numbers, have a corresponding Metal–MACiE entry, a complete list can be found at: http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/MACiE/listBy.pl?by=metal). There is significant cross-talk between Metal–MACiE and MACiE, with Metal–MACiE relying upon MACiE for the mechanism annotation, and MACiE taking the metal cofactor annotation from Metal–MACiE. We have created a detailed overview page for each metal involved in a reaction that displays the structural and chemical data for a specific metal ion on a single page. It is possible to retrieve a Metal–MACiE entry directly from MACiE, and also to go directly to the

Metal–MACiE entry for a given metal ion from the overview page within MACiE.

We now handle organic cofactors (those small molecules which are mainly composed of non-metal atoms) in a manner analogous to the amino acid residues in MACiE. Thus, we have annotated the function of the cofactor in the individual steps within MACiE, and these data are now displayed on the overview and step information pages as with the catalytic amino acid residues. As part of this remediation process, we have developed the CoFactor database (18) and, where appropriate, MACiE links out to CoFactor from the overview page, which describes the 27 different cofactors currently identified in detail from the perspective of the cofactors themselves, rather than the enzymes in which they function.

Structural data and displaying MACiE in 3D

In order to begin to understand how the local environment of the catalytic amino acid residues affects their function, we have added information on the protein structure. This section (accessible from the overview page under the 'Structural Overview' option on the side menu or from the 'Display structure information' in the general information section of the overview page) displays the biological unit representative crystal structure for the MACiE entry in an animated Jmol (19) applet (which is distinct from the reaction animations available for some entries) that shows the catalytic domains and catalytic species as a movie. We also identify the different catalytic sites present in the protein [from the CSA (12)].

For each catalytic residue, the residues contacting it have been calculated using HBPlus (20), and are shown, again in a Jmol applet with the display centred on the catalytic residue in focus. The contact information generated using HBPlus has also been used to create a query that allows a user to identify catalytic dyads and triads present in MACiE.

Furthermore, we describe the flexibility of the catalytic residues; this is assessed using the B factors as a crude measure of flexibility. Each residue in the representative PDB code is assigned an average B factor by taking the mean B factor of all the atoms in the residue. In order to cope with the large potential variation in the average B factors and to report these data in a consistent manner, the normalized B Factor (a value between 0 and 1) within the protein structure is created by ordering the average B factors in increasing size and then dividing the ranked position by the total number of residues (21). Both the average B factor and the ranked value are displayed. This section also describes the relative solvent accessibility (RSA) of the residue. This is calculated using NACCESS (22) and is shown as a percentage. Both the B factor and RSA have been added to update the analysis previously performed on a much smaller sub-set of enzymes (21). Finally, this section includes information on the number of hydrogen bond acceptor and donor contacts to the catalytic residue.

Other new annotations

Each reaction now has a reversibility tag added to the overall reaction, which makes no inference on the biological reversibility of the reaction. This reversibility is determined automatically and depends on whether one or more steps are annotated as being unknown, irreversible, or reversible. If one or more steps are annotated with the 'unknown' reversibility tag, then the overall reaction is annotated with an unknown reversibility, irrespective of what annotations the other steps have. If one or more steps are annotated with the irreversible tag, then the overall reaction is listed as irreversible, otherwise (i.e. if all steps are annotated as reversible) the overall reaction is listed as reversible.

We have also manually added a brief, textual description of the events of a reaction step. This is displayed from the entry overview page and above the image of the step's reaction on the step page.

Furthermore, we have automated the annotation of CATH domains, based upon the latest release of CATH (v 3.4.0) (23) and the links to both EzCatDB and the SFLD.

NEW METHODS FOR CHARACTERIZING AND COMPARING ENZYME MECHANISMS

MACiE is unique in containing detailed information not only on the overall reaction being performed by an enzyme, but also in the step-wise mechanism and the catalytic residues and cofactors involved in that transformation. The criterion for inclusion into MACiE is that the enzyme is distinct at some level of one or more of these aspects (mechanism, overall reaction or catalytic machinery). In order to define the similarity between enzyme reactions we thus first define similarity (calculated using a Tanimoto similarity score) for each of these three aspects separately, and then combine them to get an 'overall' entry similarity.

Defining similarity

Catalytic machinery similarity. The catalytic machinery that is carrying out the reaction is defined for the purposes of this measure as the catalytic residues and those residues binding the metal cofactor ions (to include those cases where there are only metal ions acting in the mechanism). We do not currently include the metal and organic cofactors themselves due to the fact that they are often not present in the representative crystal structure used for the 3D superimposition. The simplest method to compare this machinery is to consider the complement of the catalytic amino acid residues. However, due to the variation in the number of amino acid residues annotated as catalytic (from no amino acid residues in M0204 up to 13 in M0143 with the average entry containing only four) a simple fingerprint, in which each amino acid residue type is considered independently and counted, can produce skewed results. In order to compensate for this, we also compare the 3D coordinates of the catalytic machinery by performing a superimposition of the residues using IsoCleft (24). The final similarity is calculated by combining both the complement and superimposition measures in a 9:1 ratio.

Overall reaction similarity. MACiE contains the manual annotation of the bonds formed, cleaved and 'changed in order' for the overall and step reactions, and we have turned this annotation into a weighted (i.e. we count both the number and type of bond changed) bond change fingerprint. We have created two types of fingerprint, one that is direction-dependent (i.e. it is important that we know the C–O bond is formed), and another that is essentially direction-independent (i.e. we don't distinguish the exact nature of the bond change, just that the C–O bond is modified during the reaction). At this point stereochemistry is only annotated at the overall reaction level.

The fingerprints describing the bond changes in the overall reaction can then be compared between entries to give an estimation of overall reaction similarity. We currently do not include any measure of the substrate/product similarity, as this information is encoded in the EC number to some extent and it is interesting to observe the cases where very different EC numbers result in almost identical bond change profiles, or cases where similar EC numbers contain very different bond change profiles independent of the substrate/product similarity.

Mechanism similarity. While the similarity of the overall reactions is relatively trivial to calculate, the similarity of the 'mechanism' is more difficult. In order to simply capture the similarity between two entries at the step level, we consider the 'mechanism' as the sum of all the bond changes involved in all the steps, which we call the 'composite bond change' fingerprint. We use this, rather than the more complicated approaches used previously (25–27), as this calculation can be performed quickly on the fly, and also effectively hides differences in how annotators have marked up the reaction, e.g. an elimination followed by a proton transfer happening in two successive steps in one entry and in a concerted manner in another, and reaction sequence timings, e.g. two reactions occur in parallel in the biological system but are annotated as occurring in sequence in MACiE. In the following, when we refer to composite reaction similarity, it is this measure to which we are referring.

Defining the 'overall' entry similarity. Each fingerprint thus created can be compared using a Tanimoto similarity score for continuous variables (28), which may take a value between 0 and 1, where 0 indicates no bits in common and 1 indicates that the two fingerprints are identical. The final similarity between two entries can thus be calculated according to the following formula, in which the mechanism is considered the more important, followed by the catalytic machinery and finally the overall reaction chemistry occurring:

(0.65 * 'composite reaction') + (0.25 * 'catalytic machinery') + (0.10 * 'overall bond change')

The weights chosen are arbitrary and designed to define the similarity based on mostly the composite reaction information, but that are also informed by the catalytic machinery and overall reaction. However, each of the measures of similarity can also be investigated individually, and all four measures are displayed on the comparative overview pages.

Exploring the data in MACiE

In order to examine the differences between such sets of entries, we have developed the dataset overview pages, which display a comparative analysis of the data available within MACiE for all the entries in the set. This includes an overview of the CATH domains annotated, the number of steps involved, the catalytic machinery and overall reactions as well as the composite reaction similarity and involvement of cofactors.

Each entry now includes a section detailing sequence homologues to the current MACiE entry using the homologues as determined by the CSA [the same as previously reported (8)] and also now using a non-iterative search [using SSEARCH (29)] for a stricter definition of homology (see Supplementary Data for more detail). Furthermore, this section includes details on other MACiE entries with the same EC number (identical to the fourth level) and CATH domains where entries have at least one catalytic CATH domain in common. We also offer the option to view all similar reactions using the overall reaction bond change similarity and the composite reaction similarity, which is available from the side bar menu. Where there are similar entries at the EC or CATH domain level the similarity at the composite reaction and catalytic machinery level is shown and there is the option to compare two reactions, or to view the dataset comparison (where there are three or more entries available).

All entries in MACiE now also include links to view similar reactions from the overview page (for the overall reaction and composite bond change perspectives) and step details page (for the reaction steps). In all cases, only reactions with a Tanimoto similarity score of greater than or equal to a specific cut-off are shown. In the case of the individual reaction fingerprint, this cut off is 0.75, in the case of the composite reaction fingerprint, this cut-off is 0.65. These cut-off values are somewhat arbitrary and have been chosen to show the most similar reactions only. The cut-off value is one of the parameters of the Perl-CGI display script, and so can be altered in the HTML address of the results page by the user.

In the following subsections, specific examples are used to highlight some of the new features available for the comparative overview of sets of entries.

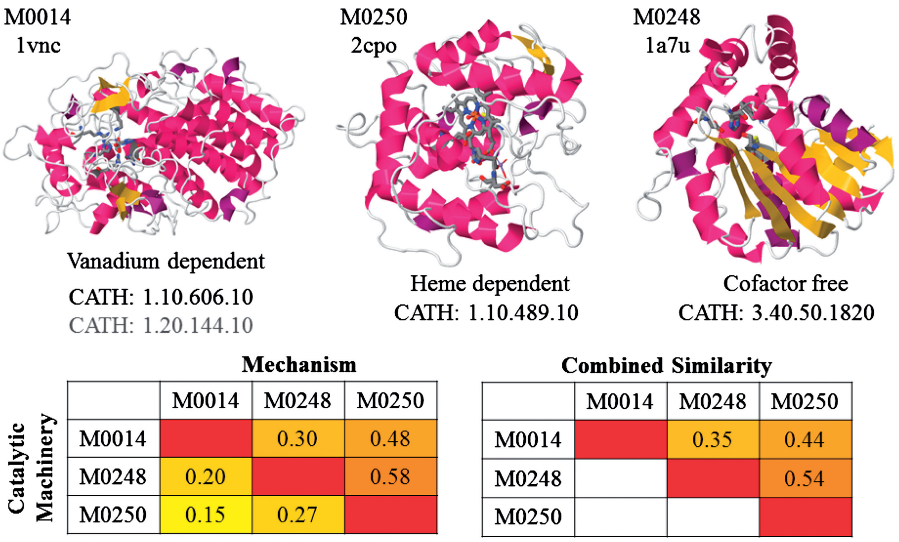
The Diversity within an EC number—the chloroperoxidases (EC:1.11.1.10). Recently (30) we investigated the number of evolutionary families present in each EC number, and found that on average each EC number had emerged approximately twice independently. Thus, there is potentially a great deal of mechanistic variability within a single EC number. While some of this variability might be related to substrate specificity for those EC numbers that are somewhat generic (e.g. EC 2.7.11.1), there are also cases where the mechanism and catalytic machinery are obviously very different. One such example is the chloroperoxidases (EC 1.11.1.10), for which there are three MACiE entries (M0014, M0248 and M0250), representing three evolutionarily unrelated families.

For this set of entries, the dataset overview pages do not display the overall reaction analysis as all these are identical, the coverage of the EC classification and the mechanisms, some of which is shown in Figure 2.

In the MACiE entries for EC 1.11.1.10 the exact method of producing the hypohalous acid (the common reactive intermediate) from a halide and hydrogen peroxide is different in all three cases. Each enzyme utilizes different catalytic CATH domains and different catalytic machinery, both in terms of amino acid residues and cofactors. These differences are reflected in the composite bond change fingerprints which fall in a relatively wide range (0.3–0.58), despite the overall reactions being identical.

Table 1 shows a selection of homologues to the entry M0248 (one of the chloroperoxidases in MACiE, UniProtKB accession 031168) within UniProtKB. The protein sequence used is taken from the PDB code used as the representative in MACiE (1a7u) and the sequence is fully annotated with the catalytic residues, their location of function and activity, the results of which can highlight where changes in the residues annotated might be related to a change in EC number and hence protein function.

The diversity within a catalytic motif—entries in MACiE containing a catalytic triad. One of the new searches added to MACiE (among several other new searches described in Supplementary section S.2 of the Supplementary Data) allows the user to search for



Mechanism

	M0014	M0248	M0250
M0014		0.30	0.48
M0248	0.20		0.58
M0250	0.15	0.27	

Combined Similarity

	M0014	M0248	M0250
M0014		0.35	0.44
M0248			0.54
M0250			

Figure 2. Similarity of the chloroperoxidase entries in MACiE. The top panel shows the 3D crystal structures with the catalytic residues and cofactors shown in ball and stick notation [images created using Jmol (20)]. The catalytic CATH domains are shown in bold text, the non-catalytic CATH domains in grey. The bottom panel shows the similarity matrices generated for these entries.

Table 1. Example results from the sequence homology for M0248

Enzyme information		Sequence similarity			Catalytic residue conservation				
UniProtKB accession	EC number	Expectation value	Percentage similarity	Chain length	32 %F	98 *S	99 %M	228 &D	257 *H
O31168	1.11.1.10	1.7e-126	100.0	277	F	S	M	D	H
P29715		7.8e-126	99.3	277	F	S	M	D	H
Q55921	1.11.1.10	2.5e-74	57.8	275	F	S	M	D	H
Q52011	3.7.1.8	6.2e-10	24.0	287	G	S	M	D	H
B7VHH1	3.1.1.1	2.5e-09	26.6	278	W	S	L	D	H
Q6Q2C2	3.3.2.10	3.4e-09	34.6	133	F	D	W	–	–
Q59695	2.3.1.12	4.7e-09	30.3	267	F	S	M	D	H
O52866	3.3.2.10	6.7e-09	28.5	221	W	D	W	–	–
P26174	6.6.1.1	0.00017	26.4	276	L	S	A	D	H
Q15N09	3.1.1.1	0.00021	23.7	253	W	S	L	D	H

The final columns of the table represent the conservation of the catalytic residues, the top line is the residue number in the sequence of the representative PDB file, the second line denotes the location of function and activity (which utilizes the following symbols: % = main chain spectator, * = side chain reactant, & = side chain spectator) followed by the single letter abbreviation for the residue. Conservative mutations are shown in green text and non-conservative mutations shown in red text.

catalytic dyads and triads. These motifs are defined as two or three residues which are hydrogen bonded to one another, and are determined automatically using HBPlus. One potential application of this search might be to identify all the entries in MACiE that utilize a Ser–His–Asp triad, as described below.

There are five entries in MACiE with a Ser–His–Asp triad where at least one of the residues is annotated as being catalytic. While the majority of these entries are in the hydrolase class of enzymes (EC 3), there are examples in the oxidoreductases (the cofactor-free choloroperoxidase, EC 1.11.1.10, M0248) and lyases (hydroxynitrilase, EC 4.1.2.37, M0217). Despite the fact that all these entries contain a Ser–His–Asp triad, these enzymes perform a distinct set of overall reactions (at the bond change only level) and have different catalytic machinery profiles, as can be seen from Figure 3. The difference in catalytic machinery may be partly related to the fact that although all these enzymes have an oxyanion hole (to stabilize the covalently attached oxyanion tetrahedral intermediate), this hole is usually made up of main chain amide groups (except in the case of M0094 where the side chain of Asn104 is one of the residues making up the oxyanion hole), and the actual identity of these residues are widely different (including Met, Phe, Leu, Glu, Gly and Tyr).

Except for the lyase example (M0217) the mechanisms are similar, and indeed contain at least four identical steps; formation of the enzyme–substrate covalently attached tetrahedral intermediate, initial elimination to re-form the carbonyl group, addition of water to the covalently attached intermediate followed by cleavage of the product from the enzyme. The variation is often either in following steps (as with the chloroperoxidase) or in the substrates involved. However, in the case of hydroxynitrilase, the catalytic triad is not acting in this manner, nor does it appear to have the standard oxyanion hole, with the substrate lacking the common carbonyl group of the other reactions' reactants. Indeed, in this enzyme the serine is simply acting as a proton shuttle and not in covalent catalysis.

Overall Reaction Bond Changes

Catalytic Machinery

	M0248	M0218	M0094	M0005	M0217
M0248		0.22	0.22	0	0
M0218	0.63		1	0.33	0.14
M0094	0.41	0.39		0.33	0.14
M0005	0.44	0.40	0.61		0.14
M0217	0.45	0.40	0.38	0.38	

Mechanism

Combined Similarity

	M0248	M0218	M0094	M0005	M0217
M0248		0.83	0.75	0.62	0.26
M0218	0.71		0.87	0.75	0.29
M0094	0.61	0.76		0.87	0.25
M0005	0.51	0.62	0.75		0.27
M0217	0.28	0.30	0.27	0.28	

Figure 3. Similarity results for the Ser–His–Asp containing entries in MACiE showing the overall bond change, catalytic machinery, mechanism and combined similarity measures.

The diversity within an evolutionarily related family. Another question that we can now address is to investigate the diversity of entries relating to a family of enzymes. We have recently shown, using the phosphatidylinositol–phosphodiesterase and Ntn-type amide hydrolase families, (N. Furnham *et al.,* submitted for publication) that there is often a good deal of variability within a family of enzymes (as represented by a single CATH domain) at the overall reaction level, as well as the structural level. This variability can be analysed in terms of the overall reaction, mechanism, composite reaction and catalytic machinery using the new overview pages. We are also starting a long-term collaboration with the SFLD, a database of 'promiscuous' enzyme superfamilies, so that all reactions in that database that fulfil the criteria for inclusion into MACiE are incorporated into our dataset. Version 3 of MACiE already incorporates a total of 26 entries from the SFLD, with all 10 structurally characterized families in the crotonase superfamily already included into MACiE.

UPDATES TO MACiE WEBSITE

Version 2.0 of MACiE (8) was based on static HTML pages. We have since moved to a model in which all the pages relating to the data content of MACiE (i.e. the lists of entries by one of the EC number, PDB code, CATH code or MACiE identifier) are generated, on the fly, by Perl CGI scripts and thus are updated automatically whenever the database is updated. Other minor changes to the online content of MACiE include the addition of mouse-over descriptions of the amino acid residue functions, mechanisms and mechanism components. These descriptions are linked to the MACiE dictionaries. We have added navigation buttons to the reaction steps, to allow users to cycle through the steps. Finally, we have added in GO terms for each entry, based on the primary PDB code and the associated UniProt accession code (31).

FUTURE DEVELOPMENTS

MACiE is an actively developing resource, and we are continuously extending its coverage. As part of this, and as mentioned before, we are working closely with the SFLD to extend coverage in MACiE of evolutionarily related superfamilies. We are beginning to work towards a new data entry system, which will be online and as automated as possible and will allow the enzyme community to add data to MACiE. We are also working on allowing users to search the intermediates in the database as well as the substrates and products, not only textually (as is currently the case), but also through substructure similarity. Furthermore, we are working on ways to handle alternative mechanisms and enzyme promiscuity more robustly. Finally, we will continue to use MACiE to attempt to understand enzymes and how they function.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online: Supplementary Sections 1–3, Supplementary Figure 1 and Supplementary References [8,9,13,21,25,30].

ACKNOWLEDGEMENTS

G.L.H. would like to thank Prof. Janet Thornton and Dr John Mitchell for their useful discussions and continued support. We would also like to thank the Wellcome Trust, EMBL and IBM (G.L.H., J.D.F., S.A.R., S.T.W.). J.D.F. is part of the European Molecular Biology Laboratory studentship programme, and is also affiliated with Cambridge University Department of Biology.

FUNDING

Ministero Instruzione Università Ricerca (RBLA032ZM7, PRIN 2007M5MWM9 to C.A.); National Library of Medicine (LM04969 to W.R.P.); National Institutes of Health (R01 GM60595 to D.E.A.). Funding for open access charge: EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. McDonald,A.G., Boyce,S. and Tipton,K.F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.*, **37**, D593–D597.
2. The UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
3. Berman,H.M., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
4. Scheer,M., Grote,A., Chang,A., Schomburg,I., Munaretto,C., Rother,M., Söhngen,C., Stelzer,M., Thiele,J. and Schomburg,D. (2011) BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.*, **39**, D670–D676.
5. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
6. Wittig,U., Golebiewski,M., Kania,R., Krebs,O., Mir,S., Weidemann,A., Anstein,S., Saric,J. and Rojas,I. (2006) SABIO-RK: Integration and Curation of Reaction Kinetics Data. *Lect. Notes Bioinformatics*, **4075**, 94–103.
7. Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. and Apweiler,R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.*, **32**, D434–D437.
8. Holliday,G.L., Almonacid,D.E., Bartlett,G.J., O'Boyle,N.M., Torrance,J.W., Murray-Rust,P., Mitchell,J.B.O. and Thornton,J.M. (2007) MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic Acids Res.*, **35**, D515–D520.
9. Holliday,G.L., Bartlett,G.J., Almonacid,D.E., O'Boyle,N.M., Murray-Rust,P., Thornton,J.M. and Mitchell,J.B.O. (2005) MACiE: a database of enzyme reaction mechanisms. *Bioinformatics*, **21**, 4315–4316.
10. Pegg,S.C., Brown,S.D., Ojha,S., Seffernick,J., Meng,E.C., Morris,J.H., Chang,P.J., Huang,C.C., Ferrin,T.E. and Babbitt,P.C. (2006) Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry*, **45**, 2545–2555.
11. Nagano,N. (2005) EzCatDB: the enzyme catalytic-mechanism database. *Nucleic Acids Res.*, **33**, D407–D412.
12. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.
13. Holliday,G.L., Mitchell,J.B.O. and Thornton,J.M. (2009) Understanding the functional roles of amino acid residues in enzyme catalysis. *J. Mol. Biol.*, **390**, 560–577.
14. Holliday,G.L., Almonacid,D.E., Mitchell,J.B.O. and Thornton,J.M. (2007) The chemistry of protein catalysis. *J. Mol. Biol.*, **372**, 1261–1277.
15. Fischer,J.D., Holliday,G.L., Rahman,S.A. and Thornton,J.M. (2010) The structures and physicochemical properties of organic cofactors in biocatalysis. *J. Mol. Biol.*, **403**, 803–824.
16. Andreini,C., Bertini,I., Cavallaro,G., Holliday,G.L. and Thornton,J.M. (2008) Metal ions in biological catalysis: from enzyme databases to general principles. *J. Biol. Inorg. Chem.*, **13**, 1205–1218.
17. Andreini,C., Bertini,I., Cavallaro,G., Holliday,G.L. and Thornton,J.M. (2009) Metal-MACiE: a database of metals involved in biological catalysis. *Bioinformatics*, **25**, 2088–2089.
18. Fischer,J.D., Holliday,G.L. and Thornton,J.M. (2010) The CoFactor database: Organic cofactors in enzyme catalysis. *Bioinformatics*, **26**, 2496–2497.
19. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/ (January 2011, date last accessed).
20. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
21. Bartlett,G.J., Porter,C.T., Borkakoti,N. and Thornton,J.M. (2002) Analysis of catalytic residues in enzyme active sites. *J Mol. Biol.*, **324**, 105–121.
22. Hubbard,S.J. and Thornton,J.M. (1993) *NACCESS, Computer Program*. Department of Biochemistry and Molecular Biology, University College London.
23. Cuff,A.L., Sillitoe,I., Lewis,T., Clegg,A.B., Rentzsch,R., Furnham,N., Pellegrini-Calace,M., Jones,D., Thornton,J. and Orengo,C.A. (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.*, **39**, D420–D426.
24. Najmanovich,R., Kurbatova,N. and Thornton,J. (2008) Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, **24**, 105–111.
25. O'Boyle,N.M., Holliday,G.L., Almonacid,D.E. and Mitchell,J.B.O. (2007) Using reaction mechanism to measure enzyme similarity. *J. Mol. Biol.*, **368**, 1484–1499.
26. Almonacid,D.E. and Babbitt,P.C. (2011) Toward mechanistic classification of enzyme functions. *Curr. Opin. Chem. Biol.*, **15**, 435–442.
27. Almonacid,D.E., Year,E.R., Mitchell,J.B.O. and Babbitt,P.C. (2010) Quantitative comparison of catalytic mechanisms and overall reactions in convergently evolved enzymes: implications for classification of enzyme function. *PLoS Comput. Biol.*, **6**, e1000700.
28. Willett,P., Barnard,J.M. and Downs,G.M. (1998) Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.*, **38**, 983–996.
29. Sierk,M.L. and Pearson,W.R. (2004) Sensitivity and selectivity in protein structure comparison. *J. Protein Sci.*, **13**, 773–785.
30. Holliday,G.L., Fischer,J.D., Mitchell,J.B.O. and Thornton,J.M. (2011) Characterising the complexity of enzymes based on their mechanisms and structures using a bio-computational analysis. *FEBS J.*, **278**, 3835–3845.
31. The Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.